Theoretical Computer Science 411 (2010) 1613–1625

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Recovering the long-range links in augmented graphs☆,☆☆

Pierre Fraigniaud[a,*], Emmanuelle Lebhar[a], Zvi Lotker[b]

[a] CNRS and University Paris Diderot, France

[b] Ben Gurion University, Israel

ARTICLE INFO

Keywords:
Small world
Doubling dimension
Bounded growth

ABSTRACT

The *augmented graph* model, as introduced in Kleinberg, STOC (2000) [23], is an appealing model for analyzing navigability in social networks. Informally, this model is defined by a pair (H, φ), where H is a graph in which inter-node distances are supposed to be easy to compute or at least easy to estimate. This graph is "augmented" by links, called *long-range* links, that are selected according to the probability distribution φ. The augmented graph model enables the analysis of *greedy routing* in augmented graphs $G \in (H, \varphi)$. In greedy routing, each intermediate node handling a message for a target t selects among all its neighbors in G the one that is the closest to t in H and forwards the message to it.

This paper addresses the problem of checking whether a given graph G is an augmented graph. It answers part of the questions raised by Kleinberg in his Problem 9 (Int. Congress of Math. 2006). More precisely, given $G \in (H, \varphi)$, we aim at extracting the base graph H and the long-range links R out of G. We prove that if H has a high clustering coefficient and H has bounded doubling dimension, then a simple local maximum likelihood algorithm enables us to partition the edges of G into two sets H' and R' such that $E(H) \subseteq H'$ and the edges in $H' \setminus E(H)$ are of small stretch, i.e., the map H is not perturbed too greatly by undetected long-range links remaining in H'. The perturbation is actually so small that we can prove that the expected performances of greedy routing in G using the distances in H' are close to the expected performances of greedy routing using the distances in H. Although this latter result may appear intuitively straightforward, since $H' \supseteq E(H)$, it is not, as we also show that routing with a map more precise than H may actually damage greedy routing significantly. Finally, we show that in the absence of a hypothesis regarding the high clustering coefficient, any local maximum likelihood algorithm extracting the long-range links can miss the detection of $\Omega(n^{5\varepsilon}/\log n)$ long-range links of stretch $\Omega(n^{1/5-\varepsilon})$ for any $0 < \varepsilon < 1/5$, and thus the map H cannot be recovered with good accuracy.

1. Introduction

Numerous papers that appeared during the last decade tend to demonstrate that several types of interaction networks share common statistical properties, encompassed under the broad terminology of *small worlds* [35–37]. These networks include the Internet, at the router level as well as at the autonomous system level, and the World Wide Web. Networks defined in various frameworks such as biology (e.g., metabolic and protein networks), sociology (e.g., movie actors

☆ This work was partially done while the third author was visiting University Paris Diderot at LIAFA. Additional supports by University Paris Diderot, COST Action 295 "DYNAMO", by ANR Project "ALADDIN", and by INRIA project "GANG" are gratefully acknowledged.

☆☆ A preliminary version of this paper appeared in the proceedings of the 15th Int. Colloquium on Structural Information and Communication Complexity (SIROCCO), Villars-sur-Ollon, Switzerland, June 17–20, 2008.

* Corresponding address: LIAFA, Université Paris Diderot, Case 7014, 75205 Paris cedex 13, France.
E-mail address: pierre.fraigniaud@liafa.jussieu.fr (P. Fraigniaud).

0304-3975/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2010.01.009

collaboration network), and linguistics (e.g., pairs of words in English texts that appear at most one word apart) also share these statistical properties [20]. Specifically, a network is said to be a small world [39] if

- it has low density, i.e., the total number of edges is "small", typically linear in the number of nodes;
- the average distance between nodes is "small", typically polylogarithmic as a function of the number of nodes; and
- the so-called *clustering* coefficient, measuring the local edge density, is "high", i.e., it is significantly higher than the clustering coefficient of Erdös–Rényi random graphs $\mathcal{G}_{n,p}$.

Other properties often shared by the aforementioned networks include:

- *scale-free* properties [6], e.g., a fat tailed shapes in the distributions of parameters such as node degree;
- limited growth of the ball sizes [2,21]; and/or
- low doubling dimension [38].

A lot remains to be done to understand why the properties listed above appear so frequently, and to design and analyze models capturing these properties. Nevertheless, there is now a common agreement on their presence in interaction networks. The reason for this agreement is that, although the statistical validity of some measurements is still under discussion [4], many tools (including the controversial Internet Traceroute) have been designed to check whether a network satisfies the aforementioned properties.

This paper addresses the problem of checking another important property shared by social networks:

- the *navigability* property.

It was indeed empirically observed that social networks not only possess small average inter-node distance, but also that short routes between any pair of nodes can be found by simple decentralized processes [9,34]. One of the first papers aiming at designing a model capturing this property is due to Kleinberg [23], where the notion of *augmented graphs* is introduced. Informally, an augmented graph aims at modeling two kinds of knowledge of distances available to the nodes: a global knowledge given by a base graph, and a local knowledge given by one extra random link added to each node. The idea is to mimic the available knowledge in social networks, where individuals share some global distance comparison tool, e.g., geographical or professional, but have also private connections, e.g., friendship, that are unknown to the other individuals. We define an *augmented graph model* as a pair (H, φ) where H is a graph, called the *base graph*, and φ is a probability distribution, referred to as an *augmenting distribution* for H. This augmenting distribution is defined as a collection of probability distributions $\{\varphi_u, u \in V(H)\}$. Every node $u \in V(H)$ is given one extra link,[1] called a *long-range link*, pointing to some node, called the *long-range contact* of u. The destination v of such a link is chosen at random with probability $\Pr\{u \to v\} = \varphi_u(v)$. If $v = u$ or v is a neighbor of u, then no link is added. In this paper, a graph $G \in (H, \varphi)$ will often be denoted by $H + R$ where H is the base graph and R is the set of long-range links resulting from the trial of φ yielding G.

An important feature of this model is that it enables to define simple but efficient decentralized routing protocols modeling the search procedure applied by social entities in Milgram's [34] and Dodd's et al. [9] experiments. In particular, *greedy routing* in (H, φ) is the oblivious routing process in which every intermediate node along a route from a source $s \in V(H)$ to a target $t \in V(H)$ chooses among all its neighbors (including its long-range contact) the one that is the closest to t *according to the distance measured in* H, and forwards to it. For this process to apply, the only "knowledge" that is supposed to be available at every node is its distances to the other nodes in the base graph H. This assumption is motivated by the fact that, if the base graph offers some "nice properties", e.g., it is embeddable in a low-dimensional metric with small distortion, then the distance function dist_H is expected to be easy to compute, or at least to approximate, locally.

Lots of effort has been made to better understand the augmented graph model. See, e.g., [1,5,7,11–15,24,29–32], and the survey [25]. Most of these works tackle the following problem: given a family of graphs $\mathcal{H}$, find a family of augmenting distributions $\{\varphi_H, H \in \mathcal{H}\}$ such that, for any $H \in \mathcal{H}$, greedy routing in (H, φ_H) performs efficiently, typically in polylog(n) expected number of steps, where $n = |V(H)|$. Kleinberg first showed that greedy routing performs in $O(\log^2 n)$ expected number of steps on any square mesh augmented with an appropriate harmonic distribution [23]. Among the works that followed Kleinberg's seminal results, an informative result due to Duchon et al. [10] states that any graph of bounded growth can be augmented so that greedy routing performs in polylog(n) expected number of steps. Slivkins [38] extended this result to graphs of bounded doubling dimension, and even doubling dimension $O(\log \log n)$. This bound on the doubling dimension is tight since [16] proved that, for any function $d(n) = \omega(\text{polylog}(n))$, there is a family of graphs of doubling dimension $d(n)$ for which any augmentation yields greedy routing performing in $\omega(\text{polylog}(n))$ expected number of steps.[2]

Despite these progresses in analyzing the augmented graph model for small worlds, the key question of its validity is still under discussion. In [25], Kleinberg raised the question of how to check that a given network is an augmented graph (Problem 9). This is a critical issue since, if long-range links are the keystone of the small world phenomenon, they should be present in social networks, and their detection should be greatly informative. This paper aims at answering part of this detection problem.

---

[1] By adding $k_u \geq 1$ long-range links to node u, for every $u \in V(H)$, instead of just one, with $\Pr(k_u = k) \sim 1/k^\alpha$ for some $\alpha > 1$, the model can also capture the scale-free property. For the sake of simplicity however, we will just assume $k_u = 1$ for every $u \in V(H)$.

[2] The notation $d(n) = \omega(f(n))$ for some functions f and d means that $d(n)/f(n)$ tends to infinity when n goes to infinity.

1.1. The reconstruction problem

This paper addresses the following reconstruction problem: given an n-node graph $G = H + R \in (H, \varphi)$, for some unknown graph H and unknown distribution φ, extract a good approximation H' of H such that greedy routing in G using distances in H' performs approximately as well as when using the distances in the "true" base graph H. More precisely, the expected number of steps of greedy routing in H' has to be the one in H up to a polylogarithmic factor. Note that, for every edge in R one extremity is the long-range contact of the other. Nevertheless, there is no a priori orientation of these edges when G is given.

To measure the quality of the approximation H' of H, we define the *stretch* of a long-range link between u and v as $\text{dist}_H(u, v)$. Then, the extracted base graph H' is considered to be of good quality if it contains H and does not contain too many long-range links of large stretch. Indeed, we want to approximate H by H' as close as possible not only for the purpose of efficient routing using the metric of H', but also because the augmented graph model assumes that distances in H are easy to compute or approximate. Therefore, the map of distances of H' should be as close as possible to the one of H.

In addition to its fundamental interest, the reconstruction problem may find important applications in network routing. In particular, if the base graph H offers enough regularity to enable distance computation using node names (or labels) of small size, then critical issues of storage and quick access to routing information such as the ones currently faced for the Internet [26,33] can be addressed. Indeed, applying greedy routing in the network using solely the distances in H may be sufficient to achieve fast routing, i.e., performing in an expected polylogarithmic number of steps.

1.2. Methodology

In statistics, one of the most used techniques is the maximum likelihood method [22]. Applied to our problem, this would lead to the extraction of the long-range links based on their probability of existence. Precisely, the method would select S as the set of the n long-range links such that

$$\Pr(G \mid S \text{ is the set of long-range links})$$

is maximum. This brute force approach however requires testing an exponential number of sets, and it requires some knowledge about the distribution φ. For instance, in [3,8], the authors assume that R is a random power law graph added on top of the base graph H. Motivated by the experimental results in [28], and the analytical results in [10,23,27,38], we consider augmenting distributions where $\varphi_u(v)$ is inversely proportional to the size of the ball of radius $\text{dist}_H(u, v)$ centered at u. We call such kind of augmenting distributions *density-based* distributions. They are the ones enabling an efficient augmentation of graphs with bounded ball growth, and, up to modifying the underlying metric by weighting nodes, of graphs with bounded doubling dimension.

Fixing a class of augmenting distributions still does not suffice for applying the maximum likelihood method because of the large number of sets. One way to overcome this difficulty is to consider every edge separately. More precisely, we consider *local* maximum likelihood methods defined as follows.

Definition 1. An algorithm $\mathcal{A}$ for recovering the base graph H from $G \in (H, \varphi)$ is a *local maximum likelihood algorithm* if and only if $\mathcal{A}$ decides whether or not an edge $e \in E(G)$ is a long-range link solely based on the value of $\Pr(G \mid e \in E(H))$.

Applying a local maximum likelihood algorithm however requires some information about the local structure of the base graph H. For instance, in [8], the base graph H is assumed to possess a clustering property characterized by a large number of edge-disjoint paths of bounded length connecting the two extremities of any edge. In [3], the clustering property is characterized by a large amount of flow that can be pushed from one extremity of an edge to the other extremity, along routes of bounded length. Motivated by the statistical evidences demonstrating that social networks are locally dense, we consider a clustering property stating that every edge participates in at least $c \cdot \log n / \log\log n$ triangles for some positive constant c. Note that this function grows very slowly, and that its output for practical values of n is essentially constant: for a network with one billion nodes, our assumption states that every edge participates to at least $6c$ triangles; and for a network with one billion billions nodes, this bound becomes $10c$. Note also that even though we focus on the number of triangles, our approach could easily be adapted to apply on many other types of local structures whose characteristics would enable distinguishing local connections from remote connections.

1.3. Our results

First, we present a simple local maximum likelihood algorithm, called EXTRACT, that, given an n-node graph $G = H + R \in (H, \varphi)$, where H has a clustering coefficient such that every edge participates in $\Omega(\log n / \log\log n)$ triangles, and φ is a density-based augmenting distribution, computes a partition (H', R') of $E(G)$. This partition satisfies $E(H) \subseteq H'$ and, for any $\beta \geq 1$, if X is the random variable counting the number of links in $R \setminus R'$ of stretch at least $\log^{\beta+1} n$, then $\Pr\{X > \log^{2\beta+1}(n)\} \leq 1/n$ whenever the maximum degree Δ of H satisfies $\Delta = O(\log^{\beta} n)$. That is, Algorithm EXTRACT is able to almost perfectly reconstruct the map H of G, up to long-range links of polylogarithmic stretch. It is worth mentioning that Algorithm EXTRACT runs in time close to linear in $|E(G)|$, and thus is applicable to large graphs with few edges, which is typically the case of small world networks.

Our main positive result (Theorem 1) is that if in addition H has bounded growth, then greedy routing in G using the distances in H' performs in polylog(n) expected number of steps between any pair. This result is crucial in the sense that Algorithm EXTRACT is able to approximate the base graph H and the set R of long-range links accurately enough so that greedy routing performs efficiently. In fact, we prove that the expected slow down of greedy routing in G using the distances in H' compared to greedy routing in (H, φ) is only polylog(n). Although this latter result may appear intuitively straightforward since $H' \supseteq E(H)$, we prove that routing with a map more precise than H may actually damage greedy routing performances significantly.

We also show how these results can be generalized to the case of graphs with bounded doubling dimension.

Finally, Theorem 3 proves that the clustering coefficient plays a crucial role for extracting the long-range links of an augmented graph using local maximum likelihood algorithms. We prove that *any* local maximum likelihood algorithm extracting the long-range links in some augmented graph with low clustering coefficient fails. In fact, this is true even in the case of cycles augmented using the harmonic distribution, that is even in the case of basic graphs at the kernel of the theory of augmented graphs [23]. We prove that any local maximum likelihood algorithm applied to the harmonically augmented cycle fails to detect $\Omega(n^{5\varepsilon}/\log n)$ of the long-range links of length $\Omega(n^{1/5-\varepsilon})$ for any $0 < \varepsilon < 1/5$.

2. Extracting the long-range links

In this section, we first focus on the task of extracting the long-range links from an augmented graph $G = H + R \in (H, \varphi)$ without knowing H. The efficiency of our extraction algorithm in terms of greedy routing performances will be analyzed in the next section. As will be shown in Section 4, extracting the long-range links from an augmented graph is difficult to achieve in the absence of a priori assumptions on the base graph H and on the augmenting distribution φ. Before presenting the main result of the section we thus present the assumptions made on H and φ.

The clustering coefficient of a graph H aims at measuring the probability that two distinct neighboring nodes u, v of a node w are neighbors. Several similar formal definitions of the clustering coefficient appear in the literature. In this paper, we use the following definition. For any node u of a graph H, let $N_H(u)$ denote the neighborhood of u, i.e., the set of all neighbors of u in H.

Definition 2. An n-node graph H has clustering $c \in [0, 1]$ if and only if c is the smallest real such that, for any edge $\{u, v\} \in E(H)$,

$$\frac{|N_H(u) \cap N_H(v)|}{n} \geq c.$$

For instance, according to Definition 2, each edge of a random graph $G \in \mathcal{G}_{n,p}$ with $p \simeq \frac{\log n}{n}$ has expected clustering $1/n^2$ up to polylogarithmic factors. In our results, motivated by the fact that interaction networks have a clustering coefficient much larger than uniform random graphs, we consider graphs in (H, φ) for which the clustering coefficient of H is slightly more that $1/n$, that is every edge participates in $\Omega(\log n/\log\log n)$ triangles.

We also focus on augmenting distributions that are known to be efficient ways to augment graphs of bounded growth (or bounded doubling dimension) [10,23,38]. For any node u of a graph H, and any $r > 0$, let $B_H(u, r)$ denote the ball centered at u of radius r in H, i.e., $B_H(u, r) = \{v \in V(G) \mid \text{dist}_H(u, v) \leq r\}$.

Definition 3. An augmenting distribution φ of a graph H is *density-based* if and only if $\varphi_u(u) = 0$, and for every two distinct nodes u and v of H,

$$\varphi_u(v) = \frac{1}{Z_u} \frac{1}{|B_H(u, \text{dist}_H(u, v))|}$$

where $Z_u = \sum_{w \neq u} 1/|B_H(u, \text{dist}_H(u, w))|$ is the normalizing coefficient.

Density-based distributions are motivated by their kernel place in the theory of augmented graphs, as well as by experimental studies in social networks. Indeed, density-based distributions applied to graphs of bounded growth roughly give a probability $1/k$ for a node u to have its long-range contact at distance k, which distributes the long-range links equivalently over all scales of distances, and thus yields efficient greedy routing. In addition, Liben-Nowell et al. [28] showed that in some social networks, two-third of the friendships are actually geographically distributed this way: the probability of befriending a particular person is inversely proportional to the number of closer people.

Notation. According to the previous discussion, for any $\beta \geq 1$, we consider the family $\mathcal{M}(n, \beta)$ of n-node density-based augmented graph models (H, φ) where H has clustering $c = \Omega(\frac{\log n}{n \log\log n})$ and maximum degree $\Delta \leq \gamma \log^\beta n$ for some constant $\gamma > 0$.

We describe below a simple algorithm, called EXTRACT, that, given an n-node graph G and a real $c \in [0, 1]$, computes a partition (H', R') of the edges of G. This simple algorithm will be proved quite efficient for reconstructing a good approximation of the base graph H and a good approximation the long-range links of a graph $G \in (H, \varphi)$ when H has high clustering and φ is density-based.

Algorithm EXTRACT:
Input: a graph G, $c \in [0, 1]$;
  $R' \leftarrow \emptyset$;
  For every $\{u, v\} \in E(G)$ **do**
    If $\frac{1}{n} |N_G(u) \cap N_G(v)| < c$ **then** $R' \leftarrow R' \cup \{u, v\}$;
  $H' \leftarrow E(G) \setminus R'$;
Output: (H', R').

Note that the time complexity of Algorithm EXTRACT is $O(\sum_{u \in V(G)} (\deg_G(u))^2)$, i.e., close to $|E(G)|$ for graphs of constant average degree. More accurate outputs could be obtained by iterating the algorithm using the test $\frac{1}{n} |N_{H'}(u) \cap N_{H'}(v)| < c$ until H' stabilizes. However, this would significantly increase the time complexity of the algorithm without significantly improving the quality of the computed decomposition (H', R'). The main quantifiable gain of iterating Algorithm EXTRACT would only be that H' would be of clustering c, and would be maximal for this property. Finally, note also that Algorithm EXTRACT involves local computations, and therefore could be implemented in a distributed manner.

The result hereafter summarizes the main features of Algorithm EXTRACT.

Lemma 1. *Let $(H, \varphi) \in \mathcal{M}(n, \beta)$, and $G \in (H, \varphi)$. Let c be the clustering coefficient of H. Assume $G = H + R$. Then Algorithm* EXTRACT *with input (G, c) returns a partition (H', R') of $E(G)$ such that $E(H) \subseteq H'$, and:*

$$\Pr(X > \log^{2\beta+1} n) = O\Big(\frac{1}{n}\Big),$$

where X is the random variable counting the number of links in $R \setminus R'$ of stretch at least $\log^{\beta+1} n$.

Proof. Since H has clustering c, for any edge $\{u, v\}$ in $E(H)$, $\frac{1}{n} |N_H(u) \cup N_H(v)| \geq c$, and therefore $\{u, v\}$ is not included in R' in Algorithm EXTRACT. Hence, $E(H) \subseteq H'$. For the purpose of upper bounding $\Pr(X > \log^{2\beta+1} n)$, we first lower bound Z_u, for any $u \in G$. We have for any $u \in G$, $Z_u \geq \deg_H(u)/(\deg_H(u) + 1) \geq 1/2$.

Let $\mathcal{S} \subseteq R$ be the set of long-range links that are of stretch at least $\log^{\beta+1} n$. We say that an edge $\{u, v\} \in R$ *survives* if and only if it belongs to H'. For each edge $e \in \mathcal{S}$, let X_e be the random variable equal to one if e survives and 0 otherwise, when R is the set of random links chosen according to φ.

Let $e = \{u, v\} \in \mathcal{S}$. For e to be surviving in H', it requires that u and v have at least $c \cdot n$ neighbors in common in G. If w is a common neighbor of u and v in G, then, since $\mathrm{dist}_H(u, v) \geq \log n > 2$, at least one of the two edges $\{w, u\}$ or $\{w, v\}$ has to belong to R. Note that u and v can only have one common neighbor w such that both of these edges are in R because we add at most one long-range link to every node, and $\{u, v\} \in \mathcal{S}$. Thus, there must be at least $c \cdot n - 1$ common neighbors w for which exactly one of the edges $\{w, u\}$ or $\{w, v\}$ is in R. The following claim upper bounds the probability of this event.

Claim 1. $\Pr\{X_e = 1\} \leq 1/n$ *where* $e = \{u, v\} \in \mathcal{S}$.

Proof. Let $w \in V(H)$ and assume that $\{w, v\} \in E(H)$. Since $\mathrm{dist}_H(u, w) \geq \mathrm{dist}_H(u, v) - 1 \geq \log^{\beta+1} n - 1$, and $\mathrm{dist}_H(v, w) = 1$, we get that $B_H(w, \mathrm{dist}_H(w, u))$ contains at least $|N_H(v)| + \log^{\beta+1} n - 2$ nodes. Therefore, the probability that u is the long-range contact of w is at most:

$$\frac{1}{Z_{\min}} \cdot \frac{1}{|N_H(v)| + \mathrm{dist}_H(u, w) - 2} \leq \frac{2}{|N_H(v)| + \log^{\beta+1} n - 2}$$

where $Z_{\min} = \min_u Z_u$.

The probability that u and v have at least $c \cdot n$ neighbors in common in G is at most the probability that there are k_1 of the nodes $w \in N_H(v)$ such that the long-range contact of w is u and k_2 nodes $w \in N_H(u)$ such that the long-range contact w is v, with $k_1 + k_2 \geq c \cdot n - 1$. Using the previous upper bound on the probability of each of these events, we get that $\Pr\{X_e = 1\}$, i.e. the probability for e to survive, is at most:

$$\sum_{k_1, k_2 \geq 0,\, k_1 + k_2 \geq c \cdot n - 1} \binom{|N_H(v)|}{k_1} \binom{|N_H(u)|}{k_2} \prod_{j=1}^{k_1} \frac{2}{|N_H(v)| + \log^{\beta+1} n - 2} \prod_{i=1}^{k_2} \frac{2}{|N_H(u)| + \log^{\beta+1} n - 2}$$

$$\leq \frac{1}{[(\log^{\beta+1} n - 2)/(2N)]^{c \cdot n - 1}} \sum_{k_1, k_2 \geq 0,\, k_1 + k_2 \geq c \cdot n - 1} \binom{|N_H(v)|}{k_1} \binom{|N_H(u)|}{k_2} \frac{1}{|N_H(v)|^{k_1}} \frac{1}{|N_H(u)|^{k_2}},$$

where $N = \max\{|N_H(u)|, |N_H(v)|\}$. Since the maximum degree is $\Delta \leq \gamma \log^{\beta} n$ and $N \leq \Delta$, we have $((\log^{\beta+1} n - 2)/(2N))^{-1} \leq 4\gamma / \log n$. Moreover, for any $a \in \mathbb{N}$, since $a!/(a-b)! \leq a^b$, we have $\binom{a}{b} \frac{1}{a^b} \leq \frac{1}{b!}$. Finally, we get that $\Pr\{X_e = 1\}$ is at most:

$$\frac{1}{(\log n/(4\gamma))^{c \cdot n - 1}} \sum_{k_1, k_2 \geq 0,\, k_1 + k_2 \geq c \cdot n - 1} \frac{1}{k_1! k_2!} \leq \frac{1}{(\log n/(4\gamma))^{c \cdot n - 1}} \sum_{i \geq c \cdot n - 1} \frac{2^i}{i!} = \frac{O(1)}{(\log n/(4\gamma))^{c \cdot n - 1}} \leq \frac{1}{n},$$

for n large enough, since $c = \Omega(\frac{\log n}{n \log \log n})$. This completes the proof of the claim. □

To compute the probability that at most $\log^{2\beta+1} n$ edges of $\mathcal{S}$ survive in total, we use virtual random variables that dominate the variables X_e, $e \in R$, in order to bypass the dependencies between the X_e. Let us associate to each $e \in \mathcal{S}$ a

random variable Y_e equal to 1 with probability $1/n$ and 0 otherwise. By definition, Y_e dominates X_e for each $e \in \mathcal{S}$ and the Y_e are independently and identically distributed. Note that, the fact that some long-range link e survives affects the survival at most Δ^2 other long-range links of R, namely, all the potential long-range links between $N_H(u)$ and $N_H(v)$. Therefore the probability that k links of $\mathcal{S}$ survive is at most the probability that k/Δ^2 of the variables Y_e are equal to one. In particular we have: $\Pr\{\sum_{e\in\mathcal{S}} X_e > \log^{2\beta+1} n\} \leq \Pr\{\sum_{e\in\mathcal{S}} Y_e > \log^{2\beta+1} n/\Delta^2\}$. Using Chernoff's inequality, we have the following claim.

Claim 2. $\Pr\{\sum_{e\in\mathcal{S}} Y_e > \log^{2\beta+1} n/\Delta^2\} \leq 1/n$.

Proof. The variables Y_e, $e \in \mathcal{S}$, are i.i.d., and $\mathbb{E}\{\sum_{e\in\mathcal{S}} Y_e\} = |\mathcal{S}|/n \leq 1$. From Chernoff's inequality, we get, for any $\delta > 0$:

$$\Pr\left\{\sum_{e\in\mathcal{S}} Y_e > (1+\delta)\cdot\mathbb{E}\left\{\sum_{e\in\mathcal{S}} Y_e\right\}\right\} < \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{\mathbb{E}\{\sum_{e\in\mathcal{S}} Y_e\}}$$
$$= n^{-\left((1+\delta)\log(1+\delta)-\delta\log e\right)\frac{\mathbb{E}\{\sum_{e\in\mathcal{S}} Y_e\}}{\log n}}$$

Let $(1+\delta) = \frac{\log^{2\beta+1} n}{\Delta^2}\cdot\frac{1}{\mathbb{E}\{\sum_{e\in\mathcal{S}} Y_e\}}$. Note that $(1+\delta) \geq \log n/\mathbb{E}\{\sum_{e\in\mathcal{S}} Y_e\} \geq \log n$, then $\delta\log e \leq (\delta+1)\log(\delta+1)/2$ for $n \geq n_1$ for some $n_1 > 0$. Therefore:

$$\Pr\left\{\sum_{e\in\mathcal{S}} Y_e > (1+\delta)\cdot\mathbb{E}\left\{\sum_{e\in\mathcal{S}} Y_e\right\}\right\} < n^{-\frac{1}{2}\log(1+\delta)} \leq n^{-\frac{1}{2}\log\log n} \leq \frac{1}{n}.$$

Finally:

$$\Pr\left\{\sum_{e\in\mathcal{S}} Y_e > \log^{2\beta+1} n/\Delta^2\right\} = \Pr\left\{\sum_{e\in\mathcal{S}} Y_e > (1+\delta)\cdot\mathbb{E}\left\{\sum_{e\in\mathcal{S}} Y_e\right\}\right\} \leq \frac{1}{n}.$$

This completes the proof of the claim. □

From Claim 2, we directly conclude that $\Pr\{\sum_{e\in\mathcal{S}} X_e > \log^{2\beta+1} n\} \leq \frac{1}{n}$. □

3. Navigability

In the previous section, we have shown that we can almost recover the base graph H of an augmented graph $G \in (H,\varphi)$: very few long-range links of large stretch remain undetected with high probability. In this section, we prove that our approximation H' of H is good enough to preserve the efficiency of greedy routing. Indeed, although it may appear counterintuitive, being aware of more links does not necessarily speed up greedy routing. In other words, using a map $H' \supseteq H$ may not yield better performance than using the map H, and actually it may even significantly damage the performances. This phenomenon occurs because the augmenting distribution φ is generally chosen to fit well with H, and this fit can be destroyed by the presence of a few more links in the map. This is illustrated by the following property.

Property 1. *There exists an n-node augmented graph model (H,φ) and a long-range link e such that, for $\Omega(n)$ source–destination pairs, the expected number of steps of greedy routing in (H,φ) is $O(\log^2 n)$, while greedy routing using distances in $H\cup\{e\}$ takes $\omega(\mathrm{polylog}(n))$ expected number of steps.*

Proof. Let $n = k^d$ with $k, d \geq 1$. We set H as the $2n$-node graph consisting in a path P of n nodes $u_1,\ldots,u_n$ connected to a d-dimensional ℓ_∞-mesh M of n nodes. Precisely, M is the n-node graph consisting of k^d nodes labeled $(x_1,\ldots,x_d)$, $x_i \in \mathbf{Z}_k$ for $1 \leq i \leq d$, where $k = n^{1/d}$. Node $(x_1,\ldots,x_d)$ of M is connected to all nodes $(x_1+a_1,\ldots,x_d+a_d)$ where $a_i \in \{-1,0,1\}$ for $1 \leq i \leq d$, and all operations are taken modulo k. Note that, by construction of M, the distance between any two nodes $x = (x_1,\ldots,x_d)$ and $y = (y_1,\ldots,y_d)$ is $\max_{1\leq i\leq d}\min\{|y_i - x_i|,\ k - |y_i - x_i|\}$. Hence, the diameter of M is $\lfloor n^{1/d}/2\rfloor$. Assume that P is augmented using the harmonic augmenting distribution h, and M is augmented using some augmenting distribution ψ. It is proved in [16] that, for any augmenting distribution ψ for M, there is a pair $s_0, t_0 \in V(M)$, with $2^{d-1}-1 \leq \mathrm{dist}_M(s_0,t_0) \leq 2^d$ such that the expected number of steps of greedy routing from s_0 to t_0 is $\Omega(2^d)$ whenever $d < \sqrt{\log n}$. Let $d = \sqrt{\log n}/2$. To construct H, we connect the extremity u_n of P to the node t_0 of M (see Fig. 1). In P, we use a slight modification $\hbar$ of the harmonic distribution h: $\hbar$ is exactly h except at node u_1 where $\hbar_{u_1}(s_0) = 1$ (i.e. for any trial of $\hbar$, the long-range contact of u_1 is s_0). Consider the augmented graph model $(H,\hbar\cup\psi)$, and set $e = \{u_1,s_0\}$.

In $(H,\hbar\cup\psi)$, greedy routing within P takes $O(\log^2 n)$ expected number of steps [23]. Let $H' = H\cup\{e\}$. We consider greedy routing using distances in H' between the two following sets:

$$S = \{u_2,\ldots,u_{\sqrt{n}}\} \quad\text{and}\quad T = \{u_{n-\sqrt{n}},\ldots,u_n\}.$$

Hence, for any $s \in S$ and $t \in T$, the shortest path from s to t in H' goes through e. Indeed, their shortest path in H is of length at least $n - 2\sqrt{n}$, while in H' it is of length at most $2\sqrt{n} + \mathrm{dist}_H(s_0,t_0) + 2 \leq 2\sqrt{n} + 2^{\sqrt{\log n}/2} + 2$ using e, which is less than $n - 2\sqrt{n}$.

Let $\mathcal{B} = B_H(u_{n-\sqrt{n}}, 2\sqrt{n} + n^{1/d})$. For any node $x \in S$, the probability that the long-range contact of x is in $\mathcal{B}$ is $O\left(\frac{1}{\sqrt{n}\cdot\log n}\right)$. Therefore, the expected number of steps required to find such a link in S is $\Omega(\sqrt{n}\cdot\log n)$ which is larger than $|S|$. As a

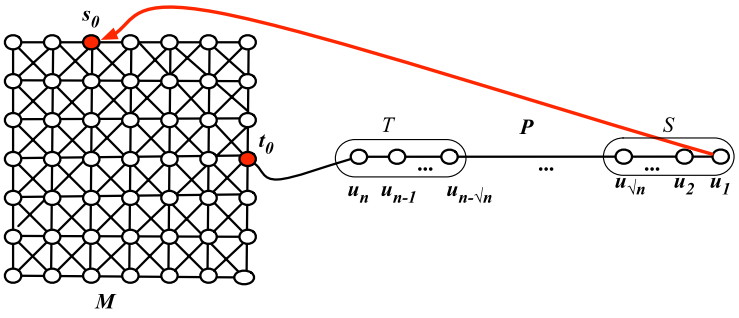

Fig. 1. Graph H in the proof of Property 1.

consequence, with constant probability, greedy routing from a node $s \in S$ to a node $t \in T$, using the distances in H', routes to u_1 and, from there to s_0. This implies that greedy routing from s to t will take at least as many steps as greedy routing from s_0 to t_0 within (M, ψ), that is $\Omega(2^{\sqrt{\log n}})$ expected number of steps, which is $\omega(\text{polylog}(n))$. □

Property 1 illustrates that being aware of some of the long-range links may slow down greedy routing dramatically, at least for some source–destination pairs. Nevertheless, we show that algorithm EXTRACT is accurate enough for the undetected long-range links not to cause too much damage. Precisely, we show that for bounded growth graphs as well as for graphs of bounded doubling dimension, greedy routing using distances in H' can slow down greedy routing in (H, φ) only by a polylogarithmic factor.

3.1. Bounded growth graphs

Definition 4. A graph G has (q_0, α)-*expansion* if and only if, for any node $u \in V(G)$, and for any $r > 0$, we have: $|B_G(u, r)| \geq q_0 \Rightarrow |B_G(u, 2r)| \leq 2^\alpha |B_G(u, r)|$. In this paper, we will set $q_0 = O(1)$, and refer to α as the *expanding dimension* of G, and to 2^α as the *growth rate* of G.

Definition 4 is inspired by Karger and Ruhl [21]. The only difference with Definition 1 in [21] is that we exponentiate the growth rate. Note that, according to Definition 4, a graph has bounded growth if and only if its expanding dimension is $O(1)$.

Theorem 1. *Let $(H, \varphi) \in \mathcal{M}(n, \beta)$ be such that H has (q_0, α)-expansion, with $q_0 = O(1)$ and $\alpha = O(1)$. Let $G \in (H, \varphi)$. Algorithm* EXTRACT *outputs (H', R') such that* (a) *$E(H) \subseteq H'$,* (b) *with high probability H' contains at most $\log^{2\beta+1} n$ links of stretch more than $\log^{\beta+1} n$, and* (c) *for any source s and target t, the expected number of steps of greedy routing in G using the metric of H' is $O(\log^{4+4\beta+(\beta+1)\alpha} n)$.*

The intuition of the proof is the following. We are given $G \in (H, \varphi)$, but Algorithm EXTRACT returns a superset H' of H. The edges in $H' \setminus H$ are undetected long-range links. Greedy routing performs according to the map H'. It is known that greedy routing according to H performs efficiently, but the undetected long-range links create a distortion of the map. Actually, the long-range links that really distort the map are those of large stretch. The standard analysis of greedy routing uses the distance to the target as potential function. For the analysis of greedy routing using the distorted map H', we use a more sophisticated potential function that incorporates the number of undetected long-range links which belong to shortest paths between the current node and the target (cf. the notion of "concerned indices" in the proof).

Proof. The fact that $E(H) \subseteq H'$ and that with high probability H' contains at most $\log^{2\beta+1} n$ links of stretch more than $\log^{\beta+1} n$ is a direct consequence of Lemma 1. Recall that $\mathcal{S} \subseteq R$ denotes the set of long-range links that are of stretch at least $\log^{\beta+1} n$. Let $H'' = H' \setminus \mathcal{S}$. By this definition, it follows that:

Claim 3. *The maximum stretch in H'' is $\log^{\beta+1} n$.*

For any $x \in V(H)$, let $L(x)$ denote the long-range contact of x. Let Z_u be the normalizing constant of the augmenting distribution at node u. We have the following claim.

Claim 4. *For any $u \in V(G)$, $Z_u \leq 2^\alpha \log n =_{\text{def}} Z_{\max}$.*

Proof. Let D be the diameter of the graph.

$$Z_u = \sum_{v \neq u} \frac{1}{|B_H(u, \text{dist}_H(u, v))|} = \sum_{r=1}^{D} \frac{|B_H(u, r)| - |B_H(u, r-1)|}{|B_H(u, r)|}$$

$$= \sum_{i=1}^{\log(D+1)} \sum_{k=2^{i-1}}^{2^i - 1} \frac{|B_H(u, k)| - |B_H(u, k-1)|}{|B_H(u, k)|}$$

$$\leq \sum_{i=1}^{\log(D+1)} \frac{1}{|B_H(u,2^{i-1})|} \sum_{k=2^{i-1}}^{2^i-1} (|B_H(u,k)| - |B_H(u,k-1)|)$$
$$\leq \sum_{i=1}^{\log(D+1)} \frac{|B_H(u,2^i-1)|}{|B_H(u,2^{i-1})|} - 1 \leq \sum_{i=1}^{\log(D+1)} (2^\alpha - 1) \leq 2^\alpha \log n.$$

This completes the proof of the claim. □

Let us analyze greedy routing in G from $s \in V(G)$ to $t \in V(G)$ using the distances in H'. Assume that $\mathcal{S} = \{\{u_1, v_1\}, \ldots, \{u_k, v_k\}\}$ is the set of the surviving long-range links (i.e. in $R \cap H'$) that have stretch more than $\log^{\beta+1} n$, v_i being the long-range contact of u_i for all $1 \leq i \leq k$. For the homogeneity of the notations, let $u_0 = v_0 = t$.

Let τ be the current step of greedy routing from s to t, and x the current node. We define the *concerned index* at step τ as the unique index j defined by:

$$j = \min_{i \in \{1,\ldots,k\}} \{i \,|\, \text{dist}_{H'}(x,t) = \text{dist}_{H''}(x,u_i) + 1 + \text{dist}_{H'}(v_i,t)\}.$$

In other words, $\{u_j, v_j\}$ is the first surviving long-range link encountered along the shortest path from x to t in H'. If there is no such index, set $j = 0$.

Claim 5. *Let x be the current node of the greedy routing, j be the concerned index at the current step, and $r > 0$. If $x \in B_{H''}(u_j, r)$, but $\text{dist}_{H''}(x,u_j) > r/2$, then the long-range contact $L(x)$ of x satisfies:*

$$\Pr\{L(x) \in B_{H''}(u_j, r/2)\} \geq \frac{1}{2^{4\alpha} \log^{1+\alpha(\beta+1)} n},$$

and if $L(x) \in B_{H''}(u_j, r/2)$ then greedy routing routes inside $B_{H''}(u_j, r/2)$ at the next step.

Proof. We have $B_H(u_j, r/2) \subseteq B_{H''}(u_j, r/2)$. Therefore, the probability for $L(x)$ to lie in $B_{H''}(u_j, r/2)$ is at least the probability to lie in $B_H(u_j, r/2)$. Moreover, the largest distance in H from x to a node in $B_H(u_j, r/2)$ is at most $\log^{\beta+1} n \cdot (3r/2)$. Indeed, in view of Claim 3, the stretch is at most $\log^{\beta+1} n$ in H''. It follows:

$$\Pr\{L(x) \in B_{H''}(u_j, r/2)\} \geq \frac{1}{Z_{\max}} \cdot \frac{|B_H(u_j, r/2)|}{|B_H(x, \log^{\beta+1} n \cdot (3r/2))|}.$$

On the other hand $B_H(x, \log^{\beta+1} n \cdot (3r/2)) \subseteq B_H(u_j, \log^{\beta+1} n \cdot (5r/2))$. And from the expanding dimension α of H we get:

$$|B_H(u_j, \log^{\beta+1} n \cdot (5r/2))| \leq 2^{\alpha(\log 5 + (\beta+1)\log\log n)} |B_H(u_j, r/2)|.$$

Finally:

$$\Pr\{L(x) \in B_{H''}(u_j, r/2)\} \geq \frac{1}{Z_{\max}} \cdot \frac{1}{2^{3\alpha} \log^{\alpha(\beta+1)} n} = \frac{1}{2^{4\alpha} \log^{1+\alpha(\beta+1)} n}.$$

Assume that the event "$L(x) \in B_{H''}(u_j, r/2)$" occurs. Suppose for the purpose of contradiction that the next step of greedy routing is a node y with $y \notin B_{H''}(u_j, r/2)$. From the greedy routing strategy, it must be that $\text{dist}_{H'}(y,t) \leq \text{dist}_{H'}(L(x), t)$. Since x has only one long-range link, y has to be a neighbor of x in H. Therefore, $\text{dist}_H(x,y) = 1 = \text{dist}_{H''}(x,y)$. Moreover, since j is the concerned index for x, y has to be on a shortest path in H'' from x to u_j (otherwise y would be further from t than x in H'). We have:

$$\text{dist}_{H'}(y,t) \geq \text{dist}_{H'}(x,t) - 1 = \text{dist}_{H''}(x,y) + \text{dist}_{H''}(y,u_j) + \text{dist}_{H'}(v_j,t)$$
$$> r/2 + 1 + \text{dist}_{H'}(v_j,t).$$

On the other hand, we have $\text{dist}_{H'}(L(x),t) \leq r/2 + 1 + \text{dist}_{H'}(v_j,t)$, and therefore we obtain that $\text{dist}_{H'}(L(x),t) < \text{dist}_{H'}(y,t)$. This is in contradiction with the greedy routing strategy, which concludes the proof of the claim. □

Claim 6. *Let x and x' be two nodes on the greedy route reached at respective steps τ and τ', $\tau < \tau'$. Assume that the concerned index at steps τ and τ' is the same, denoted by j, $j \leq k = |\mathcal{S}|$. If $x \in B_{H''}(u_j, r)$ for some $r > 0$, then $x' \in B_{H''}(u_j, r)$.*

Proof. Since $\tau' > \tau$, the greedy routing strategy enforces that $\text{dist}_{H'}(x',t) < \text{dist}_{H'}(x,t)$. On the other hand, by definition of the concerned index we have:

$$\text{dist}_{H'}(x,t) = \text{dist}_{H''}(x,u_j) + 1 + \text{dist}_{H'}(v_j,t) \leq r + 1 + \text{dist}_{H'}(v_j,t)$$
$$\text{and} \quad \text{dist}_{H'}(x',t) = \text{dist}_{H''}(x',u_j) + 1 + \text{dist}_{H'}(v_j,t),$$

therefore $\text{dist}_{H''}(x',u_j) \leq r$ and thus $x' \in B_{H''}(u_j, r)$, which completes the proof of the claim. □

For any $0 \leq i \leq \log n$, $0 \leq j \leq k$, and $\tau > 0$, let $\mathcal{E}_j^i(\tau)$ be the event: "greedy routing from s to t already entered $B_{H''}(u_j, 2^i)$ during the first τ steps". Note that, for any $0 \leq j \leq k$ and any $\tau > 0$, $\mathcal{E}_j^0(\tau) \subseteq \ldots \subseteq \mathcal{E}_j^{\log n}(\tau)$. We describe the current state of greedy routing at step τ by the event $\mathcal{E}_0^{i_0}(\tau) \cap \cdots \cap \mathcal{E}_k^{i_k}(\tau)$ where for every $0 \leq j \leq k$, $i_j = \min\{i \mid \mathcal{E}_j^i(\tau) \text{ occurs}\}$.

Note that greedy routing has reached t at step τ if and only if $\mathcal{E}_0^0(\tau)$ has occurred. Clearly, at step 0 (in s), the event $\mathcal{E}_0^{\log n}(0) \cap \mathcal{E}_1^{\log n}(0) \ldots \cap \mathcal{E}_k^{\log n}(0)$ occurs.

Claim 7. *Assume that the state of greedy routing at step τ is $\mathcal{E}_0^{i_0}(\tau) \cap \cdots \cap \mathcal{E}_k^{i_k}(\tau)$, for some $i_0, \ldots, i_k \in \{0, \ldots, \log n\}$. Then, after at most $(k+1) \cdot 2^{4\alpha} \log^{1+\alpha(\beta+1)} n$ steps in expectation, there exists an index $0 \leq \ell \leq k$ such that the state of greedy routing is $\mathcal{E}_0^{j_0}(\tau') \cap \cdots \cap \mathcal{E}_\ell^{j_\ell}(\tau') \ldots \cap \mathcal{E}_k^{j_k}(\tau')$, with $j_s \leq i_s$ for all s, $j_\ell < i_\ell$, and $\tau' > \tau$.*

Proof. At any step τ, we have for any $0 \leq j \leq k$:

$$\Pr\{\mathcal{E}_j^{i-1}(\tau+1) \mid \mathcal{E}_j^i(\tau) \text{ and } j \text{ is the concerned index at step } \tau\} \geq 1/(2^{4\alpha} \log^{1+\alpha(\beta+1)} n). \quad (1)$$

Indeed, from Claim 6, if $\mathcal{E}_j^i(\tau)$ and if j is the concerned index at the current step, then the current node x satisfies $\text{dist}_{H''}(x, u_j) \leq 2^i$ and we can apply Claim 5 which provides the above inequality.

For any fixed j, from Eq. (1), if there exist $i > 0$ and $\tau > 0$ such that $\mathcal{E}_j^i(\tau)$ occurs, then greedy routing does not perform more than $2^{4\alpha} \log^{1+\alpha(\beta+1)} n$ steps in expectation before $\mathcal{E}_j^{i-1}(\tau')$ occurs for some $\tau' > \tau$. Besides, since every step $\tau > 0$ has a concerned index, after at most $(k+1) \cdot 2^{4\alpha} \log^{1+\alpha(\beta+1)} n$ steps in expectation, there must exist one index j for which $\mathcal{E}_j^{i-1}(\tau')$ occurs for some $\tau' > 0$. This concludes the proof of the claim. □

Let X be the random variable counting the number of steps of greedy routing from s to t. As noticed before, $\mathbb{E}(X)$ is at most the expected number of steps τ to go from state $\mathcal{E}_0^{\log n}(0) \cap \mathcal{E}_1^{\log n}(0) \ldots \cap \mathcal{E}_k^{\log n}(0)$ to state $\mathcal{E}_0^0(\tau) \cap \mathcal{E}_1^{i_1}(\tau) \cdots \cap \mathcal{E}_k^{i_k}(\tau)$, for some $i_1, \ldots, i_k \in \{0, \ldots, \log n\}$. From Claim 7, we get: $\mathbb{E}(X) \leq (k+1) \log n \cdot ((k+1) \cdot 2^{4\alpha} \log^{1+\alpha(\beta+1)} n)$. And, from Lemma 1, $\Pr\{k > \log^{2\beta+1} n\} \leq 1/n$. Therefore, we have:

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X \mid k \leq \log^{2\beta+1} n) \cdot \Pr\{k \leq \log^{2\beta+1} n\} + \mathbb{E}(X \mid k > \log^{2\beta+1} n) \cdot \Pr\{k > \log^{2\beta+1} n\} \\ &\leq 2^{4\alpha} \log^{2+\alpha(\beta+1)\alpha+2(2\beta+1)} n + n \cdot (1/n) = O(\log^{4+4\beta+(\beta+1)\alpha} n). \quad \square\end{aligned}$$

Remark. Graphs of bounded expanding dimension and graphs of bounded doubling dimension are very closely related. Indeed, it can be shown that, assigning a specific weight function to a graph of bounded doubling dimension (the doubling measure of its metric), it can be made bounded growth by considering the ball sizes with nodes multiplicity corresponding to their weight [18]. Moreover, this weight function can be computed in polynomial time [19]. This allows us to extend Theorem 1 to graphs of bounded doubling dimension, up to a constant factor change in the exponent of greedy routing performances.

3.2. Graphs of bounded doubling dimension

In this section, we briefly sketch how the results for graphs of bounded expanding dimension given in Theorem 1 can be extended to graphs of bounded doubling dimension.

Definition 5. A graph G is (q_0, α)-*doubling* if and only if, for any node $u \in V(G)$, and for any $r > 0$, we have: $|B_G(u, r)| \geq q_0 \Rightarrow \exists W \subseteq V(G),\ |W| \leq 2^\alpha,\ B_G(u, 2r) \subseteq \bigcup_{w \in W} B_G(w, r)$. In this paper, we will set $q_0 = O(1)$, and refer to α as the *doubling dimension* of G.

It is easy to check that if G has (q_0, α)-expansion then G is $(q_0, 4\alpha)$-doubling (see e.g. [17]). The reverse is not true in general, except by providing to the nodes appropriate positive weights. More precisely, let us define the expansion of a node-weighted graph as in Definition 4 where the cardinality of a ball is replaced by the sum of the weights of its nodes. The following lemma is folklore (see e.g., [18,19]).

Lemma 2. *If G has (q_0, α)-expansion then G is $(q_0, 4\alpha)$-doubling. If G is (q_0, α')-doubling, then there exists a function $\mu : V(G) \to \mathbf{R}^+$ such that the node-weighted graph (G, μ) has $(q_0, 13\alpha')$-expansion. Moreover, the weights $\{\mu(u), u \in V(G)\}$ can be computed in polynomial time. We say that μ is a* doubling measure *for G.*

Using the weights introduced in Lemma 2, one can extend Definition 3: an augmenting distribution φ of a node-weighted graph $\langle H, \mathbf{w}\rangle$ is $\mathbf{w}$-density-based if and only if $\varphi_u(u) = 0$, and for every two distinct nodes u and v of H,

$$\varphi_u(v) = \frac{1}{W_u} \frac{1}{\sum_{x \in B_H(u, \text{dist}_H(u,v))} \mathbf{w}(x)}$$

where $W_u = \sum_{w \neq u} (1/\sum_{x \in B_H(u, \text{dist}_H(u,w))} \mathbf{w}(x))$ is the normalizing coefficient. Using these concepts, Theorem 1 can easily be extended to the following.

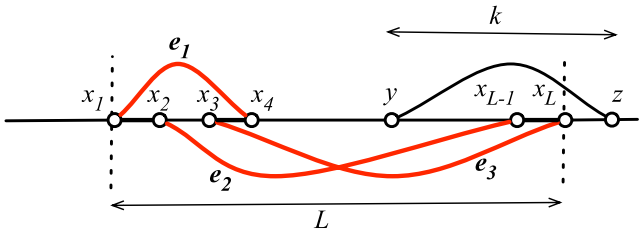

Fig. 2. Configuration of the links e_1, e_2 and e_3 of $\mathcal{E}_I$ in the proof of Theorem 3.

Theorem 2. *Let $G = (H, \varphi)$ be such that* (a) *H has clustering $c = \Omega(\frac{\log n}{n \log\log n})$,* (b) *$H$ is (q_0, α)-doubling $q_0 = O(1)$,* (c) *h has maximum degree $\Delta \leq \gamma \log^\beta n$ for some $\beta \geq 1$ and $\gamma > 1$, and (d) φ is* **w***-density-based, where* **w** *is a doubling measure for H. Algorithm* EXTRACT *outputs a partition (H', R') of $E(G)$ such that $E(H) \subseteq H'$ and for any source–target pair $(s, t) \in V(G) \times V(G)$, the expected number of steps of greedy routing in G using the distance metric of H' is $O(\log^{4+4\beta+13\alpha(\beta+1)} n)$.*

4. Impossibility results

Algorithm EXTRACT is an extreme case in the class of local maximum likelihood algorithms. Indeed, if $e = \{u, v\} \in E(H)$, one must have $\frac{1}{n}|N_G(u) \cap N_G(v)| \geq c$. Hence, if $\frac{1}{n}|N_G(u) \cap N_G(v)| < c$, then $\Pr(G \mid e \in E(H)) = 0$, and therefore it would identify e as a long-range link. Algorithm EXTRACT only fails in the detection of few long-range links with large stretch (Lemma 1) because, for a link $e = \{u, v\}$ with large stretch, $\Pr(\frac{1}{n} \cdot |N_G(u) \cap N_G(v)| \geq c)$ is small. We show that in the absence of clustering, the number of long-range links with large stretch that are not detected can be much higher, for any local maximum likelihood algorithm.

This impossibility result even holds in the case of a $(2n+1)$-node cycle C_{2n+1} augmented using the harmonic distribution $h_u^{(n)}(v) = 1/(2H_n \cdot \mathrm{dist}_{H_n}(u, v))$, where $H_n = \sum_{i=1}^{n} \frac{1}{i}$ is the nth harmonic number, and even if the extraction algorithm is designed specifically for ring base graphs augmented with the harmonic distribution.

Note that $h^{(n)}$ is density-based, but C_{2n+1} has a clustering coefficient equal to zero. It was proved in [23] that greedy routing in $(C_{2n+1}, h^{(n)})$ performs in $O(\log^2 n)$ expected number of steps between any pair.

Theorem 3. *For any $0 < \varepsilon < 1/5$, any local maximum likelihood algorithm for recovering the base graph C_{2n+1} in $G \in (C_{2n+1}, h^{(n)})$ fails in the detection of an expected number $\Omega(n^{5\varepsilon}/\log n)$ of long-range links of stretch $\Omega(n^{1/5-\varepsilon})$.*

Proof. Let $C_{2n+1} = \{x_1, \ldots, x_{2n+1}\}$ with nodes numbered clockwise. We divide C_{2n+1} into intervals of length L and we consider a specific configuration of long-range links on each of these intervals. More precisely, let $I = \{x_1, \ldots, x_L\}$ be an interval of length L in $G \in (C_{2n+1}, h^{(n)})$, and let $\mathcal{E}_I$ be the event = "G contains the six long-range links (x_1, x_4), (x_4, x_1), (x_2, x_{L-1}), (x_{L-1}, x_2), (x_3, x_L), (x_L, x_3)" (see Fig. 2). Let us write $e_1 =_{\text{def}} \{x_1, x_4\}$, $e_2 =_{\text{def}} \{x_2, x_{L-1}\}$ and $e_3 =_{\text{def}} \{x_3, x_L\}$.

Let $\mathcal{A}$ be some local maximum likelihood algorithm for detecting the long-range links. For the sake of computing a lower bound, we even assume that $\mathcal{A}$ already knows that the base graph H is C_{2n+1} and that the distribution of the augmented links is harmonic.

Note that conditionally to $\mathcal{E}_I$, $\mathcal{A}$ can make a specific mistake, called a *swap mistake*, by returning the six long-range links as local edges of C_{2n+1}, and returning the local C_{2n+1} edges $\{x_1, x_2\}$, $\{x_3, x_4\}$, and $\{x_{L-1}, x_L\}$ as six long-range links created by $h^{(n)}$. Conditionally to $\mathcal{E}_I$, counting the number of times $\mathcal{A}$ makes the swap mistake on I is a lower bound on the number of possible mistakes it does on I. Note that the swap mistake induces a modification of the distances perceived in C_{2n+1} of at most 2. For instance, the distance in C_{2n+1} from y to z in Fig. 2 is k, but it would appear as being $k+2$ if $\mathcal{A}$ does the swap mistake because the local edge $\{x_{L-1}x_L\}$ in the ring is replaced by the path $(e_2, \{x_2, x_3\}, e_3)$ of length 3. The key of the proof is to show that, when L is large, this modification is too tiny to be detectable in expectation by any local maximum likelihood algorithm. We have the following claim:

Claim 8. *Let e be a link in $G \in (C_{2n+1}, h^{(n)})$, and let $\mathcal{B}$ be a local maximum likelihood algorithm which systematically decides that*

$$e \in E(C_{2n+1}) \quad \text{if } \Pr\{G \mid e \in E(C_{2n+1})\} > \Pr\{G \mid e \notin E(C_{2n+1})\},$$
$$e \notin E(C_{2n+1}) \quad \text{if } \Pr\{G \mid e \in E(C_{2n+1})\} < \Pr\{G \mid e \notin E(C_{2n+1})\},$$
uniform random choice otherwise.

Then, the expected number of mistakes of $\mathcal{B}$ on e is at most the expected number of mistakes of any local maximum likelihood algorithm on e.

Proof. Let $\mathcal{B}'$ be some local maximum likelihood algorithm, and let $\alpha \in [0, 1]$ such that $\mathcal{B}'$ decides $e \in E(C_{2n+1})$ with probability α and $e \notin E(C_{2n+1})$ with probability $1 - \alpha$. $\mathcal{B}'$ makes a mistake on e if $e \in E(C_{2n+1})$ while it decides $e \notin E(C_{2n+1})$ or vice versa. In expectation, such a mistake occurs

$$\alpha \cdot \Pr\{G \mid e \notin E(C_{2n+1})\} + (1 - \alpha) \cdot \Pr\{G \mid e \in E(C_{2n+1})\} \text{ times.}$$

If $\Pr\{G \mid e \in E(C_{2n+1})\} > \Pr\{G \mid e \notin E(C_{2n+1})\}$, this number is strictly greater than $\Pr\{G \mid e \notin E(C_{2n+1})\}$. But in this case, $\mathcal{B}$ makes $\Pr\{G \mid e \notin E(C_{2n+1})\}$ mistakes on e in expectation. Similarly, if $\Pr\{G \mid e \in E(C_{2n+1})\} < \Pr\{G \mid e \notin E(C_{2n+1})\}$, $\mathcal{B}$ makes strictly less mistakes in expectation. Finally, if $\Pr\{G \mid e \in E(C_{2n+1})\} = \Pr\{G \mid e \notin E(C_{2n+1})\}$, the expected number of mistakes of $\mathcal{B}'$ is one while the one of $\mathcal{B}$ is 1/2. We conclude that the expected number of mistakes of $\mathcal{B}'$ is larger than the one of $\mathcal{B}$ on e. This completes the proof of the claim. □

From Claim 8, since we compute a lower bound on the expected number of mistakes of $\mathcal{A}$, we can assume that $\mathcal{A}$ always decides $e \in E(C_{2n+1})$ if $\Pr\{G \mid e \in E(C_{2n+1})\} > \Pr\{G \mid e \notin E(C_{2n+1})\}$ and $e \notin E(C_{2n+1})$ otherwise. In case of equality, $\mathcal{A}$ choses uniformly at random between the two possibilities.

We give extra power to $\mathcal{A}$ and assume that $\mathcal{A}$ even knows all local links in I except $\{x_1, x_2\}$, $\{x_3, x_4\}$, $\{x_{L-1}, x_L\}$. Since $\mathcal{A}$ knows that $H = C_{2n+1}$, it knows that the degree in H is 2. Hence, if $\mathcal{A}$ decides $e_1 \notin H$, then $\{e_2, e_3\} \notin H$ from degree considerations (indeed, there is exactly one outgoing long-range link at each node). Therefore

$$\Pr\{G \mid e_1 \notin C_{2n+1}\} = \Pr\{G \mid \{e_1, e_2, e_3\} \cap E(C_{2n+1}) = \emptyset\}.$$

Let Ω_I be the probability space describing the set of the $L-3$ other long-range links outgoing from I. A configuration $C \in \Omega_I$ can be written as $C = \left\{\{(\ell^o_5, \sigma^o_5), \mathcal{L}_5\}, \ldots, \{(\ell^o_{L-2}, \sigma^o_{L-2}), \mathcal{L}_{L-2}\}\right\}$ where for each $5 \le i \le L-2$:

- $\ell^o_i \in \{0, \ldots, n\}$ is the length of the long-range link of x_i,
- $\sigma^o_i \in \{-1, +1\}$ is the *direction* of the long-range link: it is equal to $+1$ if the link goes clockwise, and -1 otherwise,
- and $\mathcal{L}_i = \{(\ell^1_i, \sigma^1_i), \ldots, (\ell^{p_i}_i, \sigma^{p_i}_i)\}$ is the list of the lengths and directions of the $p_i \ge 0$ incoming long-range links arriving at x_i.

Note that, for any long-range link with both its extremities between x_5 and x_{L-2}, its probability of existence is unchanged whether the edges e_1, e_2, e_3 belong to C_{2n+1} or not. On the other hand, any long-range link of a node x_i, for $5 \le i \le L-2$, that has length $\ell^o_i > L - i$ and direction $\sigma^o_i = +1$ has probability $1/(2H_n(\ell^o_i + 2))$ to exist if e_1, e_2 and e_3 are in C_{2n+1}, and probability $1/(2H_n\ell^o_i)$ if e_1, e_2 and e_3 are not in C_{2n+1}. Therefore, the probability of existence of the long-range link of x_i is greater when e_1, e_2 and e_3 are not in C_{2n+1}, which is the event $\mathcal{E}_I$. Symmetrically, if the direction is $\sigma^o_i = -1$, and $\ell^o_i > i$, the probability of existence of x_i's long-range link is $1/(2H_n\ell_i)$ if e_1, e_2 and e_3 are in C_{2n+1}, and $1/(2H_n(\ell^o_i + 2))$ otherwise: the probability of existence is lower conditionally to $\mathcal{E}_I$ than to $\neg\mathcal{E}_I$. Informally, we deduce from these observations that, for any two configurations $C, \tilde{C} \in \Omega_I$ that are "symmetric" with respect to the middle of $\{x_5, \ldots, x_{L-2}\}$, $\mathcal{A}$ has to make a swap mistake on one of them. The idea of the proof is therefore to group such "symmetric" configurations in pairs in order to lower bound the expected number of swap mistakes by 1/2 on each pair.

More formally, we say that two configurations $C = \left\{\{(\ell^o_i, \sigma^o_i), \mathcal{L}_i\}, 5 \le i \le L-2\right\}$ and $\tilde{C} = \left\{\{(\tilde{\ell}^o_i, \tilde{\sigma}^o_i), \mathcal{L}'_i\}, 5 \le i \le L-2\right\}$ are *symmetric* if and only if:

1. the long-range contacts of outgoing long-range links outgoing from C or $\tilde{C}$ are not in $\{x_2, x_3, x_L, x_{L+1}\}$,
2. none of the origins of the long-range contacts ingoing in C or $\tilde{C}$ is x_{L+1},
3. for all $0 \le j \le L-7$, $(\ell^o_{5+j}, \sigma^o_{5+j}) = (\tilde{\ell}^o_{L-2-j}, -\tilde{\sigma}^o_{L-2+j})$, and $(\ell^m_{5+j}, \sigma^m_{5+j}) = (\tilde{\ell}^m_{L-2-j}, -\tilde{\sigma}^m_{L-2+j})$ for any $(\ell^m_{5+j}, \sigma^m_{5+j}) \in \mathcal{L}_{5+j}$.

Note that, because of conditions 1 and 2, not all the configurations in Ω_I can be *symmetrized*. Let Ω^s_I be the set of configuration of Ω_I that can be symmetrized. Let X_I be the random variable counting the number of swap mistakes done by $\mathcal{A}$ on I. We have:

$$\mathbb{E}(X_I \mid \mathcal{E}_I) = \sum_{C \in \Omega_I} \mathbb{E}(X_I \mid \mathcal{E}_I \text{ and } C) \cdot \Pr\{C\} \ge \sum_{C \in \Omega^s_I} \frac{1}{2} \cdot \Pr\{C\},$$

since $\mathcal{A}$ makes at least one swap mistake for two symmetric configurations C and $\tilde{C}$ in Ω^s_I in expectation. It remains to evaluate $\sum_{C \in \Omega^s_I} \Pr\{C\}$. A configuration $C = \left\{\{(\ell^o_i, \sigma^o_i), \mathcal{L}_i\}, 5 \le i \le L-2\right\}$ can be symmetrized if and only if: 1) for all $5 \le i \le L-2$, $\ell^o_i \notin \{i-2, i-3, L-i, L-i+1\}$, and 2) the long-range contact of x_{L+1} is not in $\{x_5, \ldots, x_{L-2}\}$. We get the following claim.

Claim 9. $\sum_{C \in \Omega^s_I} \Pr\{C\} \ge \mathrm{e}^{-(2\log L)/H_n}$.

Proof.

$$\begin{aligned}\sum_{C \in \Omega^s_I} \Pr\{C\} &\ge \left(1 - \frac{H_{L-3}}{2H_n}\right) \prod_{5 \le i \le L-2} \left(1 - \frac{1}{2H_n}\left(\frac{1}{i-2} + \frac{1}{i-3} + \frac{1}{L-i} + \frac{1}{L-i+1}\right)\right) \\ &\ge \left(1 - \frac{\log(L-3)}{\log n}\right) \prod_{5 \le i \le L-2} \left(1 - \frac{1}{H_n}\left(\frac{1}{i-2} + \frac{1}{L-i}\right)\right)\end{aligned}$$

Then, for n large enough,

$$\ln\left(\sum_{C\in\Omega_I^s}\Pr\{C\}\right) \geq -\frac{1}{2}\frac{\log(L-3)}{\log n} - \frac{1}{2}\frac{1}{H_n}\sum_{5\leq i\leq L-2}\left(\frac{1}{i-2}+\frac{1}{L-i}\right)$$
$$\geq -\frac{2\log L}{H_n}.$$

This completes the proof of the claim. □

Thus, $\mathbb{E}(X_I \mid \mathcal{E}_I) \geq \frac{1}{2}e^{-2(\log L)/H_n} \geq \frac{1}{2e}$, since $L \leq n$.

Let X be the random variable counting the total number of swap mistakes of $\mathcal{A}$ on G. Let $I_1, I_2, \ldots, I_{\lfloor(2n+1)/L\rfloor}$ be the largest set of adjacent and disjoint intervals of length L on C_{2n+1}. We have:

$$\mathbb{E}(X) \geq \sum_{i=1}^{\lfloor(2n+1)/L\rfloor}\mathbb{E}(X_{I_i}) \geq \sum_{i=1}^{\lfloor(2n+1)/L\rfloor}\mathbb{E}(X_{I_i}\mid\mathcal{E}_{I_i})\cdot\Pr\mathcal{E}_{I_i} \geq \sum_{i=1}^{\lfloor(2n+1)/L\rfloor}\frac{1}{2e}\cdot\frac{1}{(2H_n)^6\cdot 3^2\cdot(L-2)^4}$$

because $1/\left((2H_n)^6\cdot 3^2\cdot(L-2)^4\right)$ is the probability of existence of the six long-range links described in $\mathcal{E}_{I_i}$. Finally:

$$\mathbb{E}(X) = \Omega\left(\frac{n}{L^5\log^6 n}\right).$$

Specifically, taking $L = n^{\frac{1}{5}-\varepsilon}$ for some $0 < \varepsilon < 1/5$, we get $\mathbb{E}(X) = \Omega(n^{5\varepsilon}/\log n)$, which means that $\mathcal{A}$ fails in detecting $\Omega(n^{5\varepsilon}/\log n)$ links of stretch $\Theta(n^{\frac{1}{5}-\varepsilon})$ in expectation, since one swap mistake of $\mathcal{A}$ on some interval means that the long-range edges e_2 and e_3 of this interval have not been detected as long-range links. □

5. Conclusion

This paper is a first attempt to demonstrate the feasibility of recovering, at least partially, the base graph H and the long-range links R of an augmented graph $G = H + R$. Our methodology assumes some a priori knowledge about the structure of the base graph (of bounded doubling dimension, and with a high clustering coefficient) and of the long-range links (resulting from a trial according to a density-based distribution). It would be interesting to check whether these hypotheses could be relaxed, and, if so, to what extent.

Acknowledgements

The authors are grateful to Dmitri Krioukov for having raised to them the question of how to extract the based graph of an augmented graph, and for having pointed out several relevant references. They are also grateful to Augustin Chaintreau and Laurent Viennot for fruitful discussions.

References

[1] I. Abraham, C. Gavoille, Object location using path separators, in: 25th ACM Symp. on Principles of Distributed Computing, PODC, 2006, pp. 188–197.
[2] I. Abraham, D. Malkhi, O. Dobzinski, LAND: Stretch $(1+\epsilon)$ locality aware networks for DHTs, in: ACM-SIAM Symposium on Discrete Algorithms, SODA, 2004.
[3] R. Andersen, F. Chung, L. Lu, Modeling the small-world phenomenon with local network flow, Internet Mathematics 2 (3) (2006) 359–385.
[4] D. Achlioptas, A. Clauset, D. Kempe, C. Moore, On the bias of Traceroute sampling, or: Power-law degree distributions in regular graphs, in: 37th ACM Symposium on Theory of Computing, STOC, 2005.
[5] J. Aspnes, Z. Diamadi, G. Shah, Fault-tolerant routing in peer-to-peer systems, in: 21st ACM Symp. on Principles of Distributed Computing, PODC, 2002, pp. 223—232.
[6] A. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[7] L. Barrière, P. Fraigniaud, E. Kranakis, D. Krizanc, Efficient routing in networks with long-range contacts, in: 15th International Symposium on Distributed Computing, DISC, in: LNCS, vol. 2180, 2001, pp. 270–284.
[8] F. Chung, L. Lu, The small world phenomenon in hybrid power law graphs, in: Lect. Notes Phys., vol. 650, 2004, pp. 89–104.
[9] P. Dodds, R. Muhamad, D. Watts, An experimental study of search in global social networks, Science 301 (5634) (2003) 827–829.
[10] P. Duchon, N. Hanusse, E. Lebhar, N. Schabanel, Could any graph be turned into a small-world? Theoretical Computer Science 355 (1) (2006) 96–103.
[11] P. Duchon, N. Hanusse, E. Lebhar, N. Schabanel, Towards small world emergence, in: 18th Annual ACM Symposium on Parallel Algorithms and Architectures, SPAA, 2006, pp. 225–232.
[12] M. Flammini, L. Moscardelli, A. Navarra, S. Perennes, Asymptotically optimal solutions for small world graphs, in: 19th International Symposium on Distributed Computing, DISC, in: LNCS, vol. 3724, 2005, pp. 414–428.
[13] P. Fraigniaud, Greedy routing in tree-decomposed graphs: A new perspective on the small-world phenomenon, in: 13th Annual European Symposium on Algorithms, ESA, 2005, pp. 791—802.
[14] P. Fraigniaud, C. Gavoille, A. Kosowski, E. Lebhar, Z. Lotker, Universal augmentation schemes for network navigability: Overcoming the $\sqrt{n}$-barrier, in: 19th Annual ACM Symposium on Parallelism in Algorithms and Architectures, SPAA, 2007.
[15] P. Fraigniaud, C. Gavoille, C. Paul, Eclecticism shrinks even small worlds, in: Proceedings of the 23rd ACM Symposium on Principles of Distributed Computing, PODC, 2004, pp. 169—178.
[16] P. Fraigniaud, E. Lebhar, Z. Lotker, A doubling dimension threshold $\Theta(\log\log n)$ for augmented graph navigability, in: 14th European Symposium on Algorithm, ESA, in: LNCS, vol. 4168, 2006, pp. 376–386.

[17] A. Gupta, R. Krauthgamer, J. Lee, Bounded geometries, fractals, and low- distortion embeddings, in: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, FOCS, 2003, pp. 534–543.
[18] J. Heinonen, Lectures on Analysis on Metric Spaces, Springer-Verlag, 2001.
[19] S. Har-Peled, M. Mendel, Fast construction of nets in low dimensional metrics, and their applications, SICOMP 35 (5) (2006) 1148–1184.
[20] A. Iamnitchi, M. Ripeanu, I. Foster, Small-world file-sharing communities, in: 23rd Joint Conference of the IEEE Computer and Communications Societies, INFOCOM, 2004, pp. 952–963.
[21] D. Karger, M. Ruhl, Finding nearest neighbors in growth-restricted metrics, in: 34th ACM Symp. on the Theory of Computing, STOC, 2002, pp. 63—66.
[22] S.M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993 (Chapter 7).
[23] J. Kleinberg, The small-world phenomenon: An algorithmic perspective, in: 32nd ACM Symp. on Theory of Computing, STOC, 2000, pp. 163–170.
[24] J. Kleinberg, Small-world phenomena and the dynamics of information, Advances in Neural Information Processing Systems (NIPS) 14 (2001).
[25] J. Kleinberg, Complex networks and decentralized search algorithm, in: Nevanlinna prize presentation at the International Congress of Mathematicians, ICM, Madrid, 2006.
[26] D. Krioukov, K. Fall, X. Yang, Compact routing on Internet-like graphs, in: 23rd Conference of the IEEE Communications Society, INFOCOM, 2004.
[27] R. Kumar, D. Liben-Nowell, A. Tomkins, Navigating low-dimensional and hierarchical population networks, in: 14th European Symposium on Algorithm, ESA, in: LNCS, vol. 4168, 2006.
[28] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social networks, Proc. of the Natl. Academy of Sciences of the USA 102/3 (2005) 11623–11628.
[29] E. Lebhar, N. Schabanel, Searching for Optimal paths in long-range contact networks, in: 31st International Colloquium on Automata, Languages and Programming, ICALP, in: LNCS, vol. 3142, 2004, pp. 894–905.
[30] G. Manku, M. Naor, U. Wieder, Know thy neighbor's neighbor: The power of lookahead in randomized P2P networks, in: 36th ACM Symp. on Theory of Computing, STOC, 2004, pp. 54–63.
[31] C. Martel, V. Nguyen, Analyzing Kleinberg's (and other) small-world models, in: 23rd ACM Symp. on Principles of Distributed Computing, PODC, 2004, pp. 179–188.
[32] C. Martel, V. Nguyen, Analyzing and characterizing small-world graphs, in: 16th ACM-SIAM Symp. on Discrete Algorithms, SODA, 2005, pp. 311–320.
[33] C. Martel, V. Nguyen, Designing networks for low weight, small routing diameter and low congestion, in: 25th Conference of the IEEE Communications Society, INFOCOM, 2006.
[34] S. Milgram, The small-world problem, Psychology Today (1967) 60–67.
[35] M. Newman, The structure and function of complex networks, SIAM Review 45 (2003) 167–256.
[36] M. Newman, A. Barabasi, D. Watts (Eds.), The Structure and Dynamics of Complex Networks, Princeton University Press, Princeton, 2006.
[37] R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet: A Statistical Physics Approach, Cambridge University Press, 2004.
[38] A. Slivkins, Distance estimation and object location via rings of neighbors, in: 24th Annual ACM Symposium on Principles of Distributed Computing, PODC, 2005, pp. 41—50.
[39] D. Watts, S. Strogatz, Collective dynamics of small-world networks, Nature 393 (1998) 440–443.